Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Modelos Dinâmicos para a produção de Café no Brasil

Gabriel Ravi de Sousa dos Santos

Orientador: Professor Leandro Tavares Correia

Brasília

2018

Gabriel Ravi de Sousa dos Santos

Modelos Dinâmicos para a produção de Café no Brasil

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Professor Leandro Tavares Correia

Brasília

2018

Gabriel Ravi de Sousa dos Santos
Modelos Dinâmicos para a produção de Café no Brasil/ Gabriel Ravi de Sousa dos Santos. – Brasília, 2018-
Orientador: Professor Leandro Tavares Correia

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação II, 2018.
1. TCC 1. 2. TCC 2. 3. Pesquisa. 4. LaTeX. 5. Motivação. 6. Dedicação.

Gabriel Ravi de Sousa dos Santos

Modelos Dinâmicos para a produção de Café no Brasil

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 02 de maio de 2018:

Professor Leandro Tavares Correia
Orientador

Antônio Eduardo Gomes
Membro da Banca

Guilherme S. Rodrigues
Membro da Banca

Brasília
2018

Este trabalho é dedicado ao meu pai Wilton, minha mãe Verilda e ao meu Professor Orientador Leandro Tavares Correia.

Agradecimentos

- Primeiramente, gostaria de dedicar este trabalho a minha família, que sempre me deu suporte, e principalmente ao meu pai Wilton que nunca desistiu de mim e sempre me deu a base para estudar e trabalhar;
- Aos meus amigos, Lo Sciuto, Dario e Cristiano pelo apoio, longas conversas de bar e paciência durante as vezes que tive que me ausentar de suas vidas pelas vezes que tive que estudar;
- A minha amiga, Gabriela Vasconcelos, por estar sempre presente, ajudar nas matérias, broncas para estudar e prestar atenção na aula e, principalmente, pela amizade e companheirismo durante toda a gradução;
- Ao meu orientador Leandro Tavares Correia, que sempre foi muito paciente com as minhas dúvidas e questionamentos, empolgado com as minhas ideias e por fazer deste trabalho algo prazeroso;
- Por fim, a Empresa Júnior ESTAT Consultoria que sempre me motivou a me esforçar cada vez mais para aprender mais em sala de aula e me motivar a aplicar a estatística também fora de aula.

"Que tal criamos uma nova regra de vida...
sempre tentar ser um pouco mais gentil que o necessário."
(Extraordinário)

Resumo

Esta monografia é sobre a aplicação da metodologia de Modelos Dinâmicos Lineares a produção de café no Brasil com base nos dados de produção de café que a CONAB (Compania Nacional de Abastecimento) forneceu desde 2002. O banco de dados conta com a produção mensal de café colhido por mês e também algumas variáveis climáticas que não foram utilizadas na monografia. A técnica utilizada é a Modelagem Dinâmica Linear (soma de um modelo Sazonal com um modelo de Tendência Linear) e foi possível obter uma estimativa para o ano de 2017 de produção condizente com a realidade. A abordagem bayesiana dos modelos dinâmicos foi escolhido porque possui a interessante capacidade de incorporar, por meio da priori, no modelo, todas as informações relevantes disponíveis: desde dados históricos, experiências concretas ou subjetivas, assim como conhecimento de fenômenos futuros.

Palavras-chave: Modelos Lineares Dinâmicos, Modelos Dinâmicos Lineares, Produção de Café, Brasil

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

Depois da água, o café é a bebida mais consumida no mundo, o Brasil faturou cerca de US$5,4 bilhões com exportações de café em 2016.

O objetivo deste projeto é estudar a produção de café (em mil sacas) no Brasil e regiões brasileiras. Desenvolver uma metodologia de análise e previsão da produção de café é extremamente importante identificar o comportamento, tendências e sazonalidade da produção a fim de conhecermos a produção por região, melhorarmos o desempenho da produção de café, e direcionarmos os planos de ação que a CONAB (Companhia Nacional de Abastecimentos) utiliza para manter a produção sempre favorável.

A CONAB (Companhia Nacional de Abastecimento) é, atualmente, a responsável pela coleta e análise dos dados de produção de café no Brasil. A companhia conta com uma equipe de economistas, engenheiros e estatísticos nos quais realizam as análises devidas todo mês, e atualmente, eles utilizam uma metodologia própria de previsão. Em junho de 2017, entramos em contato com a mesma e adquirimos o banco de dados da produção de café (em mil sacas) no Brasil desde 2002, e assim, em troca, a CONAB solicitou, uma apresentação dos resultados obtidos ao final da pesquisa.

Os modelos de séries temporais têm grande utilidade para análise de comportamento e previsão de modelos de produção, dados financeiros, dados epidemiológicos, entre outras demandas.

E dentro das séries temporais, a abordagem bayesiana dos modelos dinâmicos possui a interessante capacidade de incorporar, por meio da priori, no modelo, todas as informações relevantes disponíveis: desde dados históricos, experiências concretas ou subjetivas, assim como conhecimento de fenômenos futuros. Além de previsões rotineiras, exceções podem também ser implementadas por antecipações ou em bases retrospectivas.

Muitas séries, como sequências de DNA, preço de estoques, vibrações de pontes (Xueping Fan, 2016), séries de retorno de longa duração (Chopin, 2007) possuem tal heterogeneidade temporal, assim, a utilidade de modelos dinâmicos cujos parâmetros evoluem de maneira estocástica e tal evolução é descrita por uma estrutura Markoviana.

O modelo dinâmico geral é composto por um um sistema de equações nas quais a equação das observações explica comportamento dos dados e a variável latente é desenvolvida de acordo com o tempo e tem por finalidade estimar os parâmetros da equação das observações.

1 Objetivos

O objetivo geral deste projeto é estudar e aplicar modelos dinâmicos lineares bayesianos na produção de café do Brasil e grandes regiões brasileiras.

A CONAB (Companhia Nacional de Abastecimento) tem a responsabilidade de fazer levantamentos de safras dos principais grãos, cana-de-açúcar, produtos relacionados com a agroenergia, e principalmente café. As informações são obtidas nos principais municípios, em contatos feitos com os produtores rurais, agrônomos e técnicos de cooperativas, secretárias de agricultura, órgãos de assistência técnica e extensão rural e agentes financeiros. O trabalho é realizado com periodicidade mensal, com deslocamentos a campo bimestralmente e complementada com contatos via telefone, mensagem eletrônica ou outros meios disponíveis, para atualização dos dados.

No levantamento da safra de café, a CONAB utiliza parcerias na maioria das estados produtores e os agentes prestam as informações em nível estadual e o resultado final é de responsabilidade da CONAB. A periodicidade é quadrimestral (de quatro em quatro meses).

O banco de dados possui a densidade de café, área plantada, quantidade de sacas de café produzido por mês desde Janeiro de 2002 e algumas variáveis climáticas como temperatura e estação do ano. Além disso, existem algumas variáveis explicativas como estação do ano e temperatura por mês.

Será feito uma revisão bibliográfica de séries temporais, inferência com abordagem bayesiana e modelagem preditiva. Além disso, previamente serão testados diversos modelos de previsão como o ARIMA clássico e o modelo polinomial dinâmico.

Por fim, será aplicado o pacote *dlm* a fim de obtermos a melhor modelagem aos dados. Este trabalho será realizado com o auxílio do software estatístico: R

2 Revisão Bibliográfica

2.1 Principio da Máxima Verossimilhança

O principio da máxima verossimilhança é um dos procedimentos utilizados para obter-se estimadores. Dado uma variável aleatória X com função de probabilidade (se X for uma variável aleatória discreta) ou função densidade de probabilidade (se X for uma variável aleatória contínua) $f(x, \theta)$ e dado θ um parâmetro desconhecido. Seja uma população X, retira-se uma amostra de tamanho n, $x_1, x_2, ..., x_n$, assim, a função de verossimilhança é dada por:

$$L(\theta, x_1, ..., x_n) = f(x_1, \theta) \times f(x_2, \theta) \times , ..., f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

A função de verossimilhança de uma variável aleatória discreta X é:

$$L(\theta, x_1, ..., x_n) = p(x_1, \theta) \times p(x_2, \theta) \times , ..., p(x_n, \theta) = \prod_{i=1}^{n} p(x_i, \theta)$$

E assim, este principio é base para as estimações posteriores do relatório.

2.2 Inferência Bayesiana

A inferência bayesiana é um tipo de inferência estatística que descreve as incertezas sobre quantidades não observáveis de forma probabilística na qual possui como bases a Distribuição *a priori* (conhecimento prévio dos parâmetros) e a Verossimilhança (dados observados). O teorema de Bayes é descrito por:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

A utilização de informação a *priori* em inferência Bayesiana requer a especificação de uma distribuição a priori $\pi(\theta)$ para a quantidade de interesse θ e essa distribuição a priori deve representar o conhecimento que se tem sobre θ antes da realização do experimento. E denominamos $X_1, X_2, ..., X_n$ como uma amostra de observações previamente observadas.

Resumidamente, o método Bayesiano nos ensina como combinar *a priori* $\pi(\theta)$ de hoje com a informação que acabamos de obter $L(\theta; X_1, X_2, ..., X_n)$ para chegar *a posteriori* de hoje $\pi(\theta|X_1, X_2, ..., X_n)$.

É possível obter a distribuição a *posteriori* por meio de:

$$\pi(\theta|X_1, X_2, ..., X_n) \approx \quad \pi(\theta)L(\theta; X_1, X_2, ..., X_n)$$

O $\pi(\theta|X_1, X_2, ..., X_n)$ é a distribuição *a posteriori*, $\pi(\theta)$ é a distribuição *a priori* e, por fim, $L(\theta; X_1, X_2, ..., X_n)$ é a função de verossimilhança.

2.3 Análise de Regressão Linear

A análise de regressão é um instrumento eficaz para verificar a relação entre uma variável resposta quantitativa e uma ou mais variáveis explicativas, as quais podem ser tanto qualitativas quanto quantitativas. Essa análise é feita por meio do estudo de uma função de regressão entre as variáveis estudadas. A equação abaixo exemplifica como essa função pode ser escrita:

$$Y = \alpha + \beta X + \varepsilon$$

É evidenciado na equação acima o comportamento de uma variável dependente Y em função de uma variável X, chamada de variável independente ou explicativa. O termo β indica o quanto a medida que Y varie se uma unidade de X for variada e o coeficiente α mostra o valor esperado da variável Y se X fosse nulo. Além disso, o termo ε indica o erro associado à equação em estudo.

Uma generalização do modelo de regressão simples é o modelo de regressão múltipla, no qual são consideradas mais de uma variável independente na equação. Dessa forma, a função será dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Os coeficientes são interpretados de maneira semelhante: β_0 indica o valor de Y se todas as variáveis X_i ($i = 1, 2, \ldots, k$) forem nulas; β_i mostra a variação de Y para a variação de uma unidade na variável X_i quando todas as outras variáveis são mantidas constantes; e ε informa o erro associado à equação em estudo.

2.3.1 Coeficiente de determinação na Regressão

O coeficiente de determinação, também chamado R^2, indica o quanto da variação da variável Y é explicado pelas variáveis independentes (X_1, X_2, ..., X_k). Esse coeficiente varia entre 0 e 1, indicando em porcentagem quanto está sendo explicado pelo modelo, ou seja, quanto mais perto de 1 mais as variáveis independentes explicam sobre a variação de Y. Seu valor é obtido a partir da fórmula:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SQE}{SQT}$$

com:

- n = tamanho da amostra
- $\bar{y}$ = média amostra da variável resposta Y
- $\hat{y}_i$ = valores preditos pela regressão
- $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SQE$ = soma de quadrados explicada
- $\sum_{i=1}^{n} (y_i - \bar{y})^2 = SQT$ = soma de quadrados total

Entretanto, como a soma de quadrados explicada aumenta com a adição de uma variável ao modelo, seja essa relevante ou não para explicar a variação de Y, criou-se uma adaptação do coeficiente de determinação, chamada $R^2_{ajustado}$, o qual é dados por:

$$R^2_{ajustado} = 1 - \frac{n-1}{n-(k+1)} \left(1 - R^2\right)$$

Assim, o coeficiente é penalizado com a introdução de uma nova variável e, se essa variável não for significativamente necessária para explicar a resposta, mesmo com o aumento da soma de quadrados explicada, o $R^2_{ajustado}$ diminuirá. Portanto, é possível comparar modelos com quantidades de variáveis explicativa, podendo ser utilizado para a seleção do modelo que melhor se ajusta.

2.4 Séries Temporais

Uma série temporal é uma sequência de realizações (observações) de uma variável ao longo do tempo. Dito de outra forma, é uma sequência de pontos (dados numéricos) em ordem sucessiva, geralmente ocorrendo em intervalos uniformes.

Os principais objetivos de uma análise temporal são descrever eficientemente o comportante da série, encontrar periodicidades na série e identificar os causadores de tais comportamentos. Além disso, predizer o comportamento futuro da série por meio da construção de planos a curto, médio e longo prazo e obter o direcionamento para as tomadas de decisões.

As séries de tempo são formadas pelo conjunto de informações $\{Y_t, t\epsilon T\}$, onde Y é a variável de interesse e t conjuntos de índices. As séries são classificadas como discreta ou contínua.

Uma análise de séries eficaz é capaz de construir um modelo que condiza com a realidade dos dados que, quando feito da melhor forma possível, pode propiciar apredizagem e uma previsão precisa.

Além disso, definiremos D_t como toda a informação obtida até o tempo t, e também, assume-se D_0 como a informação determinada previamente da série.

3 Modelos Dinâmicos Lineares

Considere uma série temporal Y_t definida nos tempo t = 1,2,... T, onde cada vetor aleatório Y_t tem dimensão $r \times 1$. O Modelo Dinâmico Linear (MDL) é especificado por duas equações. Uma para a série temporal observável, denominada *equação da observações*, e outra para descrever a evolução temporal dos estados latentes θ_t, denominada *equação do sistema*. Sabe-se que as informações observadas anteriores são representadas por D_{t-1}. Observe as equações abaixo.

Equação das observações

$$Y_t = F_t'\theta_t + v_t, \quad v_t \sim N[0, V_t]$$

Equação do sistema

$$\theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim N[0, W_t]$$

A cada instante de tempo t, o modelo é caracterizado pela quadrupla F_t, G_t, V_t, W_t, onde:

- F_t é uma matriz conhecida $(n \times r)$;
- G_t é uma matriz conhecida $(n \times n)$);
- V_t é uma matriz de variâncias conhecida $(r \times r)$;
- W_t é uma matriz de variâncias conhecida $(n \times n)$.

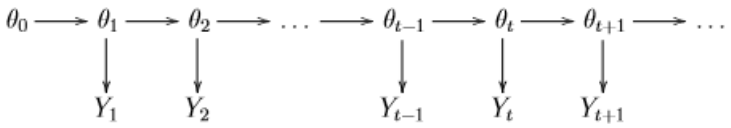
O modelo relaciona Y_t ao vetor de parâmetros do sistema θ_t $(n \times 1)$, além de especificar a evolução temporal de $\theta's$. Isso se dá através das distribuições definidas sequencialmente

$$(Y_t|\theta_t) \sim N_r[F_t'\theta_t, V_t];$$
$$(\theta_t|\theta_{t-1}) \sim N_n[G_t\theta_{t-1}, W_t]$$

A equação das observações define a distribuição de Y_t como função de θ_t. A equação do sistema supõe em uma evolução *markoviana* para o sistema. Ela é definida por uma transformação linear $G_t\theta_{t-1}$ somada a um erro aleatório com média zero.

É suposto que $(\theta_0|D_0) \sim N_n(m_0, C_0)$ e que os erros v_t e w_t são mutuamente independentes. Sabe-se que para os modelos dinâmicos, dado as informações no presente,

Figura 1 – Estrutura de dependência para o espaço de estados



o passado e o futuro são independentes. Observe a estrutura de dependência condicional do modelo dinâmico linear abaixo.

Assim, pode-se dizer que a figura acima representa bem a definição do MDL. Por fim, além disso, existem duas propriedades essenciais para os MDL's, que são:

- θ_t é uma cadeia de Markov;
- Condicionalmente a θ_t, os $Y_t s$ são independentes e Y_t depende de θ_t apenas.

3.1 Estimação do sistema e previsão

Em um modelo dinâmico linear (MDL) as variâncias V_t e W_t são conhecidas, assim, o processo de estimação é dividido em três partes:

- Evolução da Série;
- Previsão de uma nova observação
- Atualização dos parâmetros do sistema

Sabe-se que a evolução da série consiste na equação do sistema na qual utiliza a distribuição preditiva de θ_{t-1} para determinar a priori do tempo t.

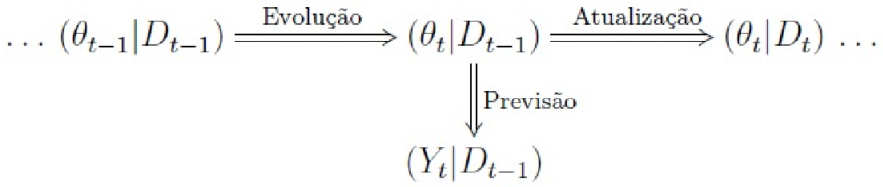
A previsão de uma nova observação é dada pela distribuição marginal de $(y_t|D_{t-1})$. Assim, obtêm-se a atualização dos parâmetros por meio da combinação da priori e a equação de verossimilhança, ambas no tempo t, por meio do teorema de Bayes.

Assim, é necessário extrair as distribuições condicionais do sistema $(\theta_s|D_t)$, onde têm-se

- $s < t$ têm-se o Processo de Suavização;
- $s = t$ têm-se o Processo de Filtragem;
- $s > t$ têm-se o Processo de previsão.

Todos os processos serão descritos nas seções em sequência.

Figura 2 – Análise sequencial do processo de estimação do modelo dinâmico linear



3.2 Processo de Filtragem

Na medida que se tem acesso a uma nova observação, aplica-se a filtragem a fim de obter estimativas para a distribuição *aposteriori* do processo latente iterativamente. Ao considerar a distribuição inicial de θ_t e as propriedades descritas previamente, o processo de filtragem pode ser descrito por:

(i) **Distribuição a *posteriori* no tempo** $t-1$

Cada média m_{t-1} e uma variância C_{t-1} são obtidos continuamente, onde

$$(\theta_t|D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$$

(ii) **Distribuição a *priori* no tempo** t

Dispõe-se $a_t = m_{t-1}$ e $R_t = C_{t-1} + W_t$, então:

$$(\theta_t|D_{t-1}) \sim N[a_t, R_t]$$

(iii) **Distribuição preditiva um passo a frente**

Dispõe-se $f_t = a_t$ e $Q_t = R_t + V_t$, então:

$$(Y_t|D_{t-1}) \sim N[f_t, Q_t]$$

(iv) **Distribuição a *posteriori* no tempo** t Têm se tal distribuição:

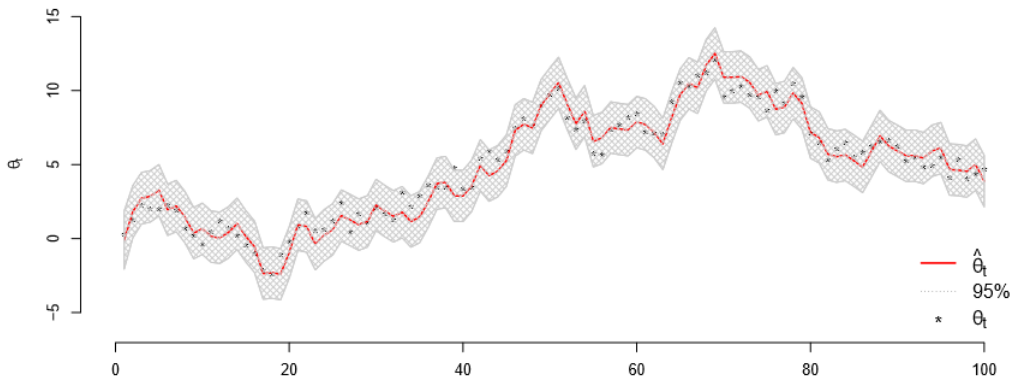
$$(\theta_t|D_t) \sim N[m_t, C_t]$$

No qual $m_t = a_t + A_t e_t$, $C_t = R_t + A_t^2 Q_t$, onde $A_t = R_t/Q_t$ e $e_t = Y_t - f_t$.

As equações apresentadas fornecem o procedimento recursivo no processo de estimação dos parâmetros do modelo. As equações acima de atualização da distribuição de probabilidade *a posteriori* são conhecidas como **Filtro de Kalman**.

As demonstrações acima são obtidas por meio do Teorema de Bayes e propriedades da distribuição Normal, conferir West e Harrison (1997) para maiores detalhes.

Figura 3 – Estimativas dos estados latentes



E também, como exemplo, a figura 3 exibe o processo de filtragem descrito e aplicado ao modelo com $V_t = 1$ e $W_t = 1$

3.3 Processo de Suavização

A suavização ou análise retrospectiva, resumidamente, dado as observações passadas, toma-se um tempo $t-k$ que toda a série observada é verificada com o objetivo de reavaliar a inferência feita pelo experimento sequencial. A reavaliação acontece por conta da informações colhidas após o período de interesse, assim, agregando mais valor a série estimada.

É possível notar que:

$$E(\theta_t|D_{t+k}) = E(E(\theta_t|\theta_{t+1}, D_{t+k})|D_{t+k})$$

Assim, por meio das propriedades descritas, pode-se dizer que $y_{t+1}, y_{t+2}, ..., y_{t+k}$ são condicionalmente independentes a θ_t, dado θ_{t+1}, portanto:

$$p(\theta_{t+1}, \theta_t|D_{t+k}) = p(\theta_{t+1}, \theta_t|D_t)$$

Dessa forma, a média suavizada têm o seguinte resultado:

$$E(\theta_t|D_{t+k}) = E(E(\theta_t|\theta_{t+1}, D_t)|D_{t+k})$$

E, utilizando as mesmas propriedades têm-se que a a nova variância estimada será:

$$Var(\theta_t|D_{t+k}) = Var(E(\theta_t|(\theta_{t+1}, D_t)|D_{t+k})) + E(Var(\theta_t|(\theta_{t+1}, D_t)|D_{t+k})$$

3.4 Processo de Previsão

O processo de previsão é de suma importância no dia-a-dia de diversas profissões. A previsão de valores futuros de uma variável qualquer, baseada no referido mecanismo de processamento, torna-se pouco confiável em situações complexas ou quando se trata de grandes decisões pois existem muitas dificuldades e presença de muitas variáveis subjetivas nos estudos.

Quando deseja-se estimar a série k-passos a frente condicionados às informações de todas as observações e do estado latente no tempo t, utiliza-se:

$$p(\theta_{t+k}|D_t) = \int p(\theta_{t+k}|\theta_{t+k-1})p(\theta_{t+k-1}|D_t)d\theta_{t+k-1}$$

E assim, obtêm-se as observações as estimativas para as observações futuras Y_{t+k} por meio da nova equação das observações:

$$p(Y_{t+k}|D_t) = \int p(Y_{t+k}|\theta_{t+k})p(\theta_{t+k}|D_t)d\theta_{t+k}$$

3.5 Fator de Desconto

O tratamento descrito acima não pode ser realizado de forma analítica, utiliza-se o *Fator de Desconto*, o qual existe para suprir a incerteza relativa a variância dos erros de evolução W_t. Quando V_t é conhecido, W_t pode ser especificado por essa técnica chamada Fator de Desconto.

O valor das informações da série decrescem ao longo tempo e, assim, a equação do sistema controla está queda por meio do acréscimo de incerteza. Sabe-se que:

$$\theta_t = G_t\theta_{t-1} + \omega_t, \quad \omega_t \sim N[0, V_t]$$

Assim,

$$V(\theta_{t-1}|D_{t-1}) = C_t \quad e \quad V(G_t\theta_{t-1}|D_{t-1}) = G_tC_{t-1}G_t' = P_t$$

Segue que:

$$R_t = V(\theta_t|D_{t-1}) = P_t + W_t$$

Então, ao definir o fator de desconto δ como a proporção da informação perdida entre os períodos t e $t-1$ e pode-se assumir um R_t tal que $R_t = P_t/\delta$, então, W_t pode ser descrito por:

$$W_t = R_t - P_t = P_t(\delta^{-1} - 1)$$

onde $0 < \delta \leq 1$.

O parâmetro δ é o fator de desconto, e é ele que "desconta"a matriz P_t que, é determinística na evolução do estado da matriz R_t. Se $\delta = 1$, assume-se que $W_t = 0$ e não existe perda de informação na passagem de θ_{t-1} para θ_t, assim $Var(\theta_t|D_t) = Var(G_t\theta_{t-1}|D_{t-1}) = P_t$, ou seja, o modelo permanece estático.

Quanto menor o fator de desconto maior será a perda de informação. E então, costuma-se assumir valores para δ acima de 0,9 para sistemas com poucas variações intensas e valores abaixo de 0,8 para sistemas que contém muita incerteza, ou seja, sistemas que possuem intervalos de predição muito largos. Quanto maior o valor de δ, mais suave são as mudanças do modelo.

3.6 Abordagem Bayesiana - Procedimentos Online e Offline de Estimação

Dado modelos de espaço de estados de duas séries temporais θ_t e Y_t, ambos com $0 < t < \infty$, satisfazendo as duas suposições abaixo:

- θ é uma cadeia de Markov;
- Y_t's são independentes entre si e dependem condicionalmente de θ_t e do vetor de parâmetros ψ.

Agora vamos observar os casos em que não conhecemos a matriz W_t e casos que não conhecemos a matriz V_t.

3.6.1 Especificação de W_t por fatores de desconto

É possível especificar as matrizes de W_t pois elas são parte da construção da distribuição *a priori*, porém, é complicado encontrar uma calibragem adequadra para a mesma e, caso esteja equivocada, pode gerar consequências ao ajuste do modelo. Sabe-se que W_t é uma matriz simétrica e quanto maiores os elementos da diagonal W_t, maior a incerteza sobre a evolução dos dados pois a quantidade de informação perdida da amostra é perdida na evolução do parâmetro θ_{t-1} para θ_t.

Nesta abordagem, torna-se inviável a estimação de W_t e assim, como solução, deseja-se especificar esta matriz da forma mais adequada, e assim, como descrito na metodologia acima, Harrison e Scott (1965) propuseram o Fator de Desconto. Lembrando que:

$$W_t = R_t - P_t = P_t(\delta^{-1} - 1)$$

onde $0 < \delta \leq 1$.

Pode-se generalizar esta metodologia a ponto que um Modelo Dinâmico Linear possa ter um Fator de desconto para cada bloco.

Além disso, a especificação de um fator de desconto para cada bloco agrega a consequência de que as estabilidades de cada efeito podem ser diferentes das outras e assim, pode-se entender que os componentes sazonais costumam ser mais estáveis que os componentes de tendência, e assim, é comum utilizar um fator de desconto de 0.95 para a tendência e 0.98 para a sazonalidade.

3.7 Modelos Dinâmicos Lineares com V_t desconhecido

Habitualmente, as variâncias observacionais, V_t, dos modelos dinâmicos lineares são aplicáveis quando as matrizes de covariância V_t são estáticas, ou seja, $V_t = \mathrm{V}$. Será atribuído $\phi = V^{-1}$ sendo o parâmetro de precisão das observações. No caso univariado, a ideia principal é atribuir uma distribuição *a priori* Gama para ϕ como $p(\phi|D_0) = F(.,.)$. Preserva-se a forma fechada do filtro de Kalman para as distribuições e assim, pode-se dizer que as distribuições marginais do sistema são t-Student.

E assim, será possível encontrar que V por sua estimativa dada por $E(\Phi|D_t)$, e partir dessa decisão, será possível resolver analiticamente $p(\theta|D_t)$ por meio da equação abaixo:

$$p(\theta|D_t) = \int p(\theta_t, \phi|D_t)d\phi = \int p(\theta_t|\phi, D_t)p(\phi|D_t)d\phi$$

3.8 Métodos de Monte Carlo via Cadeias de Markov

Os métodos de Monte Carlo baseados na simulação de variáveis aleatórias de uma Cadeia de Markov, são, atualmente, a maneira usual de fazer uma análise Bayesiana dos dados, chamados de Métodos de Monte Carlo via Cadeias de Markov (MCMC). Nas próximas seções, vamos apresentar técnicas para simulação de um vetor aleatório qualquer $(y_1, y_2, ..., y_n)$ que, num contexto bayesiano, representa os parâmetros desconhecidos de um modelo. Sendo assim, os métodos de simulação que iremos apresentar são utilizados em inferência bayesiana para simular da distribuição *a posteriori*.

O objetivo desta seção é mostrar uma alternativa possível para encontrar as distribuições *a posteriori* $\pi(\theta|Y_1, Y_2, ..., Y_n)$, que por muitas vezes é desconhecida.

E assim, os principais Métodos de Monte Carloa via Cadeias de Markov (MCMC) a serem discutidos neste trabalho são:

- Amostrador de Gibs
- Metroplis-Hastings

A seguir virão algumas técnicas de como estimar a distribuição *a posteriori* $\pi(\theta|Y_1, Y_2, ..., Y_n)$ mesmo sem conhecer os parâmetros do modelo.

3.8.1 Amostrador de Gibbs

O método mais utilizado de MCMC é o Amostrador de Gibbs, que é composto pelas distribuições condicionais completas, isto é, a partir de n valores simulados das

distribuições condicionais, e assim, chegar a um resultado para a distribuição *a posteriori*. O principal problema para a implementação do amostrador de Gibbs é ser capaz de conseguir as distribuições condicionais $\pi(Y_i|Y_1, Y_2, ..., Y_n)$.

Dado um vetor aleatório não independente $(Y_1, Y_2, ..., Y_n)$ onde as variáveis aleatórias possuem relações complexas, suponha que seja possível obter a distribuição de cada uma das variáveis aleatórias do vetor descrito acima dado os valores das variáveis, e que seja também possível simular o valor das respectivas distribuições condicionais.

Dado que:

$$Y_{-i} = (y_1, ..., y_{i-1}, y_i, ..., y_T)$$

Têm-se as n distribuições condicionais:

$$f(y_1|Y_{-1}) = f(y_1|y_2, y_3, ..., y_T)$$

$$f(y_2|Y_{-2}) = f(y_2|y_1, y_3, ..., y_T)$$

$$\vdots$$

$$f(y_n|Y_{-n}) = f(y_n|y_2, y_3, ..., y_{T-1})$$

E assim é definido uma disposição incial arbitrária para o vetor Y que é descrita por:

$$y^{(0)} = (y_1^{(0)}, ..., y_{i-1}^{(0)}, y_i^{(0)}, ..., y_T^{(0)})$$

assim que constituímos o primeiro fator da matriz de valores que desejamos e assim, interativamente, os outros fatores (linhas) vão sendo completados. É necessário simular consecutivamente cada elemento do vetor Y para preencher cada uma das linhas. Observe:

$$y_1^{(i)} \sim f(y_1|y_2^{(i-1)}, y_3^{(i-1)}, y_{T-1}^{(i-1)}, ..., y_T^{(i-1)})$$

$$y_2^{(i)} \sim f(y_2|y_1^{(i)}, y_3^{(i-1)}, y_{T-1}^{(i-1)}, ..., y_T^{(i-1)})$$

$$y_3^{(i)} \sim f(y_3|y_1^{(i)}, y_2^{(i)}, y_{T-1}^{(i-1)}, ..., y_T^{(i-1)})$$

$$\vdots$$

$$y_n^{(i)} \sim f(y_T|y_1^{(i)}, y_2^{(i)}, y_4^{(i)}, ..., y_{T-1}^{(i)})$$

Conforme as interações vão acontecendo, as atualizações das distribuições também. Um novo valor de $y_1^{(i)}$ é obtido para a variáveis aleatória Y_1 e assim sucessivamente pois utilizaremos o valor de $y_1^{(i)}$ para encontrar o valor de $y_2^{(i)}$ da variável aleatória Y_2. Assim, as interações são feitas consecutivamente até n.

4 Simulação

Abaixo será registrado uma simulação de Modelos Dinâmicos Lineares. O pacote do R chamado "dlm"fornece condições integradas para Inferência Bayesiana usando Modelos Dinâmicos Lineares. O pacote contém funções que, por exemplo, elaboram o processo de suavização ou análise retrospectiva, o processo de filtragem ou até estimação pelo método de máxima verossimilhança.

4.1 Especificação do Modelo e Estimativas

Como citado, será exemplificado o modelo de tendência estável ou **modelo polinomial de primeira ordem**. Ele é composto apenas de um nível que varia segundo um passeio aleatório:

$$y_t = \mu_t + v_t, \quad v_t \sim N[0, V]$$

$$\mu_t = \mu_{t-1} + \omega_t, \quad \omega_t \sim N[0, W]$$

Segundo esse modelo, o nível permanece localmente constante, mas varia quando se considera largos períodos de tempo e a variação das observacões em torno dos níveis (medida por V) é bem maior que as variações temporais do nível ao longo do tempo (medidas por W).

Foram construídas função de estimação e previsão do modelo dinâmico para o modelo acima. As matrizes que definem o modelo são:

$$F = [1]\ G = [1]\ \theta_t = \mu_t$$

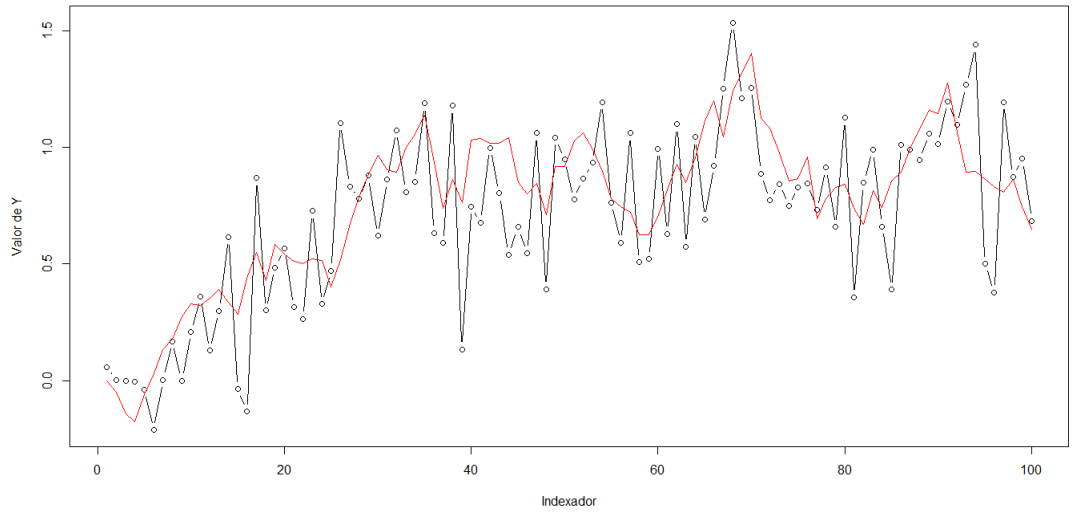
Assumimos também que as variâncias do modelo são:

$$V = [0.25]\ W = [0.1]$$

Foi construída função de estimação e previsão do modelo dinâmico para o modelo acima. Observe o gráfico com os valores simulados de $y(t)$.

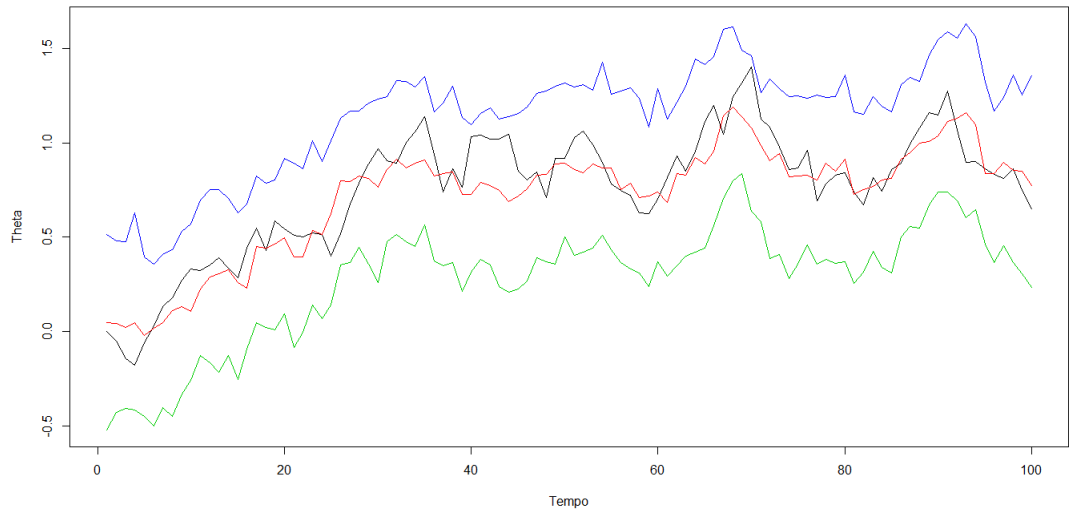
Ao observar a Figura 5 nota-se que o comportamento de y_t possui uma tendência crescente ao longo do tempo. Além disso, a linha vermelha indica a suavização da série na qual acompanha e demonstra bem a realidade da série

Figura 4 – Gráfico dos valores de y_t e θ_t



Observe abaixo os valores estimados e o intervalo de credibilidade de 95% de $\theta(t)$.

Figura 5 – Valores estimados e intervalo de credibilidade de θ



Ao observar a estimativa de θ após o processo de suavização e filtragem, nota-se que ele possui comportamento similar a y(t), e vale ressaltar que os valores estimados de θ estão próximos próximos dos valores reais do mesmo. A linha azul e a linha verde exibem o intervalo de credibilidade de 95% de θ, que no caso, se comportam de forma similar a série.

Serão especificados os modelos polinomiais, sazonais e de regressão pois são os modelos mais comuns e especiais da classe dos Modelos dinâmicos Lineares. Conforme descrito acima, o interesse desta parte é entender as funções de previsões descrevendo as

matrizes F e G.

4.2 Modelos Polinomiais

Os Modelos Dinâmicos Lineareas Polinomiais são utilizados para descrever tendências que evoluem de forma não drástica no tempo. Funções de crescimento quadrático, ou seja, modelos de baixa ordem (até 3ª ordem) podem ser utilizados para previsões de um curto ou médio prazo de tempo.

Assumindo F e G constantes, um modelo polinomial de ordem n pode ser definido por:

$$f_t(k) = E(Y_{t+k}|D_t) = a_{t0} + a_{t1}k + ... + a_{t,n-1}k^{n-1} \quad k \geq 0$$

dado que:

Assim, sabe-se que $a_{t0} + ... + a_{t,n-1}$ são funções da média a posteriori m_t, e além disso, são independentes de k.

Assim, podemos exemplificar um modelo polinomial com 4 períodos sendo:

- $y_0 = \begin{bmatrix} y_4 \\ y_3 \\ y_2 \end{bmatrix}$
- $G = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
- $W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
- $F = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$

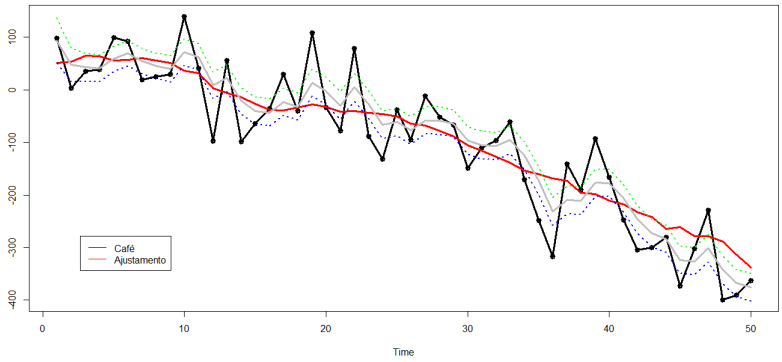
4.2.1 Exemplo de Modelo Polinomial de Segunda Ordem

- $W = \begin{pmatrix} 20 & 0 \\ 0 & 1 \end{pmatrix}$
- $V = 50$
- $F = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$
- $C_0 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$
- $m_0 = (500, 0)$

Como mencionado na metodologia, utilizou-se o pacote *dlm* para o desenvolvimento das simulações e aplicações. Além disto, para definir este modelo, foi aplicado a função *dlmModPoly*. Fixando a distribuição a *priori* sendo como $(\theta_0|D_0) \sim N_2(m_0, C_0)$ e todos os gráficos abaixo terão como média μ_t em vermelhos. E também, será possível obter os respectivos intervalos de confiança pelas linhas azuis e verdes.

Assim, utilizando o pacote *dlmFilter* é possível estimar μ_t, para $0 < t < 200$. Observe o gráfico da série filtrada abaixo:

Figura 6 – Série Simulada após Processo de Filtragem



Após isto, a fim de suavizar θ_t, será utilizado a função *dlmSmooth* abaixo:

É possível notar acima que o comportamento da série é mais contido. Agora, a fim de obter-se a série Y_t simulada um passo a frente foi utilizado a função *dlmFilter*

Figura 7 – Série Simulada após Processo de Suavização

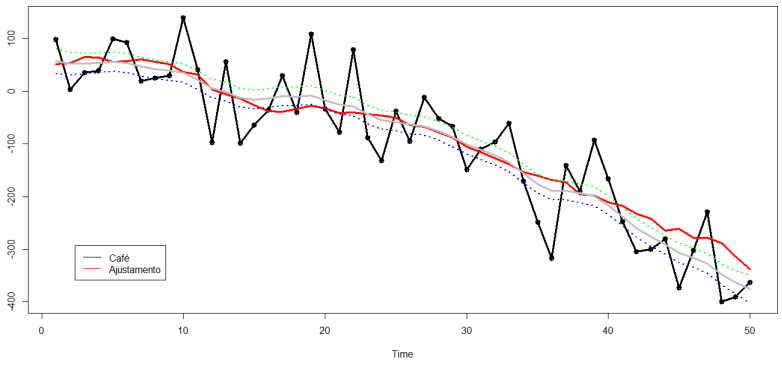
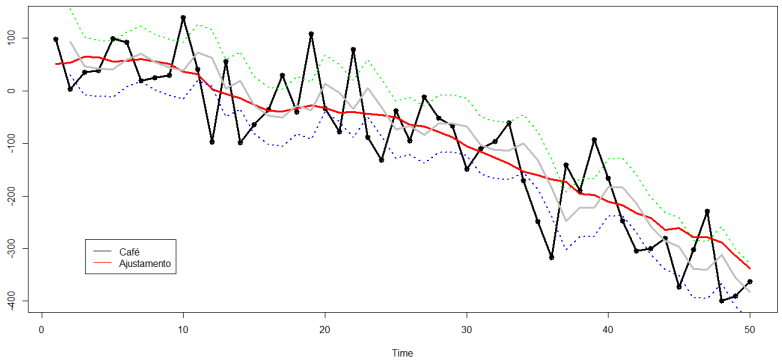


Figura 8 – Série Simulada após Processo de Previsão um passo a frente



4.3 Modelos Sazonais

Como a série da produção de café é uma série com sazonalidade bem definida, será simulada uma série com sazonalidade, e assim, aplicar os mesmos processos de suavização, filtragem e previsão um passo a frente para verificar a aplicabilidade desta técnica nos dados reais.

4.3.1 Exemplo de Modelo Sazonal

Foram utilizados referências que sejam similares aos dados da produção de café no Brasil. Os parâmetros do modelo sazonais adotados foram:

- $W = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$
- $V = 10$

- $F = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$

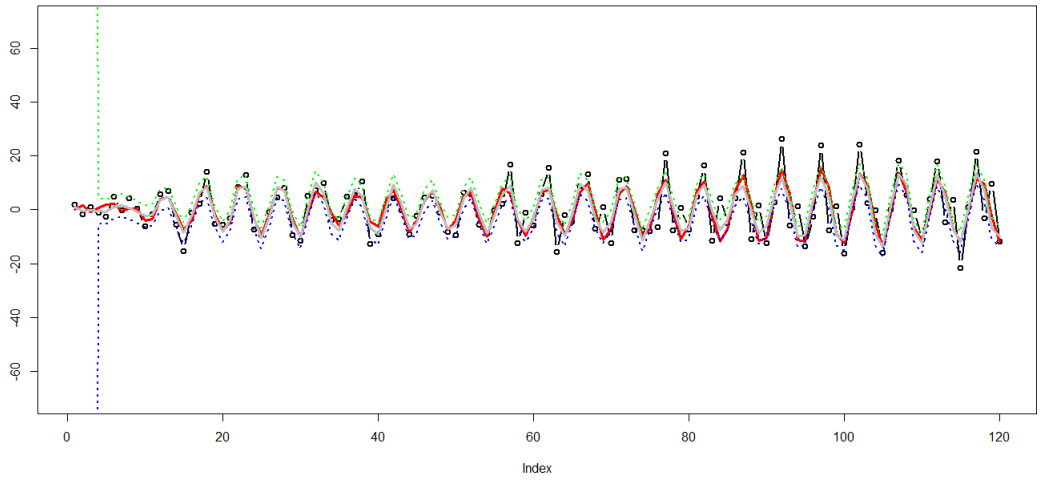
- $G = \begin{pmatrix} (1, 2\pi/5) & 0_{2x2} \\ 0_{2x2} & (1, 4\pi/5) \end{pmatrix}$

- $C_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

- $m_0 = (-0.1, 0.6, -1.1, 0.72)$

Foi gerado uma série de tamanho igual a 120 e uma distribuição *a priori*. Observe a série após a filtragem acima:

Figura 9 – Série Sazonal após Processo de Filtragem



A série acima possui 10 períodos sazonais, a linha vermelha representa as médias μ_t, a linha cinza representa a série filtrada e as linhas azuis e verdes representam os intervalos de confiança selecionados.

Após esse processo, oberseve a mesma série, porém, após o processo de suavização:

Figura 11 – Série Sazonal após Processo de Previsão

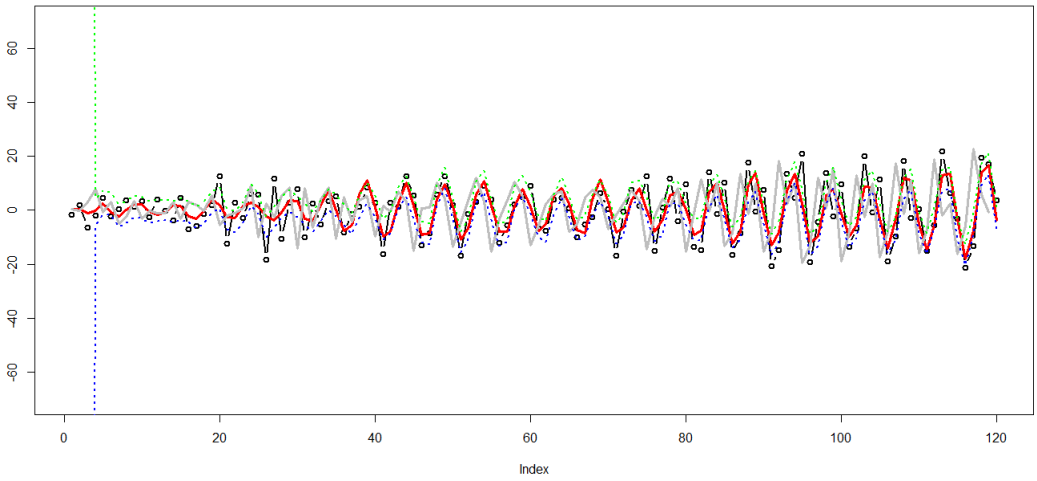
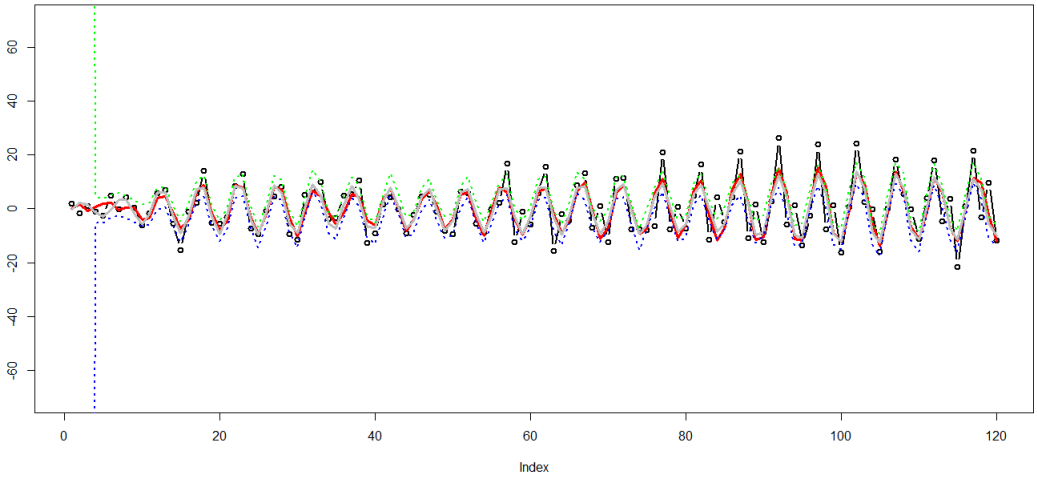


Figura 10 – Série Sazonal após Processo de Suavização



Agora, a linha cinza representa a série suavizada, assim como no modelo polinomial de segunda ordem, a série encontra-se mais contida e controlada. Agora observe série Y_t estimada um passo a frente foi utilizado o pacote *dlmFilter*.

4.4 Modelos com Parâmetros Desconhecidos

A fim de simplicar a explicação sobre as propriedades dos Modelos Dinâmicos Lineares, foi suposto que os valores da quadra F_t, G_t, V_t, W_t eram conhecidos, porém, muitos casos isso não acontece. Geralmente, W_t e V_t não são conhecidos, porém, G_t e F_t são especificados.

Nesta subseção, sem perda de generalidade, cada uma das matrizes dos modelos a serem vistos dependerão de um vetor de parâmetros desconhecidos ψ. Neste caso, a matriz dos parâmetros é estática ψ, e assim, serão utilizados diversos métodos recursivos para a estimação.

Usualmente utiliza-se para estimar o vetor ψ por **Máxima Verossimilhança**, assim, para isso, é necessário encontrar os valores que maximizem a função de verossimilhança. Observe a função de verossimilhança abaixo.

$$L(\psi) = p(y_1, y_2, ..., y_n|\psi) = \prod_{t=1}^{n} p(y_t|D_{t-1}, \psi)$$

O estimador de máxima verossimilhança de ψ é o valor que maximiza a equação acima, ou seja,

$$\hat{\psi} = argmax_{\psi} l(\psi)$$

Assim, a partir da Matriz de Fisher observada, calculada no ponto $\hat{\psi}$, extrai a matriz de covariâncias para o estimador de máxima verossimilhança de ψ. E assim, é possível utilizar este método simplismente a partir da função *dlmMLE* do pacote *dlm*.

Porém, existem fragilidades do Estimador de Máxima Verossimilhança quando estamos tratando de Modelos Dinâmicos Lineares. O processo de maximização da função de verossimilhança torna-se inviável quando o vetor ψ é muito grande pois é muito provável que a superficie pode contar diversos máximos locais ou o seu máximo não ser tão diferenciado dos outros pontos.

Como possível solução, posiciona-se a incerteza do vetor ψ por uma distribuição a priori $p(\psi)$. Para $n > 0$, assume-se que:

$$(\theta_0, \theta_1, ..., \theta_n, Y_1, Y_2, ..., X_n, \psi) \sim p(\theta_0|\psi)p(\psi) \prod_{t=1}^{n} p(y_t|\theta_t, \psi)p(\theta_t|\theta_{t-1}, \psi)$$

E assim, têm-se como observado o vetor $(y_1, y_2, ..., y_t)$ e assumindo $s = 1...$, pode-se utilizar a seguinte equação:

$$p(\theta|D_t) = \int p(\theta_s|\psi, D_t)p(\psi, D_t)d\psi$$

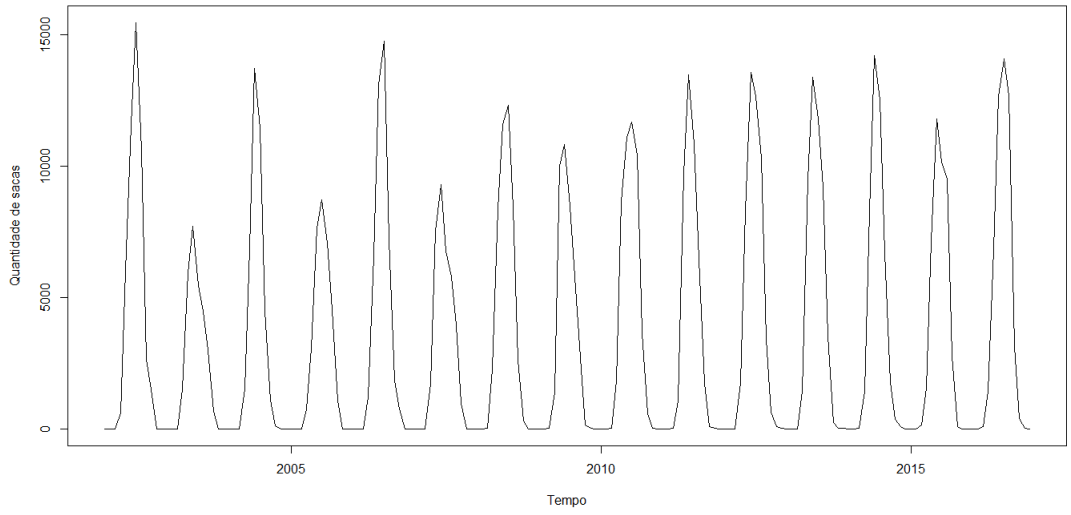
Como descrito na metodologia, em casos mais complexos, tais distribuições precisam ser aproximadas por métodos de simulção probabilistica Markov Chain Monte Carlo (MCMC), ou seja, estratégia *offline*, e existem abordagens recursivas que são ditas *online*, sendo ambas abordagens bayesianas.

5 Análise da Safra de Café no Brasil

5.1 Análise Descritiva da Safra de Café

Nesta parte iremos analisar descritivamente a safra de café por mês e ano resultando em um visão geral do comportamento da produção de café Arábico e Conilon (em mil sacas) desde o ano de 2002. O café é uma planta com sazonalidade, por isso, será possível observar que as produções nos meses de Novembro, Dezembro, Janeiro e Fevereiro não possuem produção significativa para a exportação, dado que a maior parte da produção desses meses são de autônomos e pequenos produtores específicos. Observe abaixo a série da produção de café no Brasil desde o ano de 2002.

Figura 12 – Produção de Café no Brasil(Em mil sacas) desde 2002



Ao observar a figura 12, nota-se que os anos de 2002, 2004 e 2006 são os anos que possuem a maior produção de café na série histórica, porém, nota-se que os anos com menor produção são os anos 2003, 2005 e 2008.

Ao analisar o gráfico acima, nota-se que a safra possui uma sazonalidade nítida de produção que são entre os meses de Março e Outubro. E também, nota-se que ao longo dos últimos anos, existe uma tendência crescente para a produção de café Arábico e Conilon, porém, ocorreu uma pequeno decréscimo no ano de 2015.

Observe a tabela abaixo contendo a produção de café (em mil sacas) por mês desde o ano de 2002.

Tabela 1 – Quantidade de café produzido no Brasil (Em mil sacas) desde 2002

Ano/Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
2002	0	0	0	630,24	6011,52	11586,72	15465,12	10617,12	2666,4	1502,88	0	0
2003	0	0	0	1527,46	5936,92	7723,76	5446,98	4495,92	2997,28	691,68	0	0
2004	0	0	0	1531,608	6872,6	13705,928	11428,152	4516,28	1099,616	117,816	0	0
2005	0	0	0	757,712	3096,736	7643,008	8730,16	7082,96	4546,272	1087,152	0	0
2006	0	0	0	1228,824	6599,24	13107,456	14745,888	7190,896	1865,992	773,704	0	0
2007	0	0	0	1731,408	7538,839	9306,318	6781,348	5807,431	3931,739	973,917	0	0
2008	0	0	45,9921	2299,605	8370,5622	11590,0092	12325,8828	8416,5543	2621,5497	321,9447	0	0
2009	0	0	39,469	1420,884	9985,657	10814,506	8485,835	5801,943	2723,361	157,876	39,469	0
2010	0	0	48,0948	1827,6024	8657,064	11109,8988	11687,0364	10436,5716	3655,2048	577,1376	48,0948	0
2011	0	0	43,4842	1087,105	9696,9766	13480,102	10827,5658	6479,1458	1739,368	86,9684	43,4842	0
2012	0	0	0	1778,924	8284,7032	13570,6488	12655,7736	10317,7592	3405,3688	660,7432	101,6528	50,8264
2013	0	0	0	1425,3964	9486,2588	13369,2352	11796,384	9191,3492	3538,9152	245,758	49,1516	49,1516
2014	0	0	45,3418	1450,9376	8252,2076	14191,9834	12514,3368	6710,5864	1768,3302	362,7344	90,6836	0
2015	0	0	172,94	1556,46	7177,01	11803,155	10160,225	9511,7	2767,04	86,47	0	0
2016	0	0	102,7384	1438,3376	6832,1036	12636,8232	14075,1608	12688,1924	3133,5212	410,9536	51,3692	0

Nota-se pela tabela acima que os meses de Janeiro e Fevereiro não possuem produção significativa para o Brasil, o mês de março possui produção apenas nos anos de 2008, 2009, 2010, 2011 e 2014 em diante.

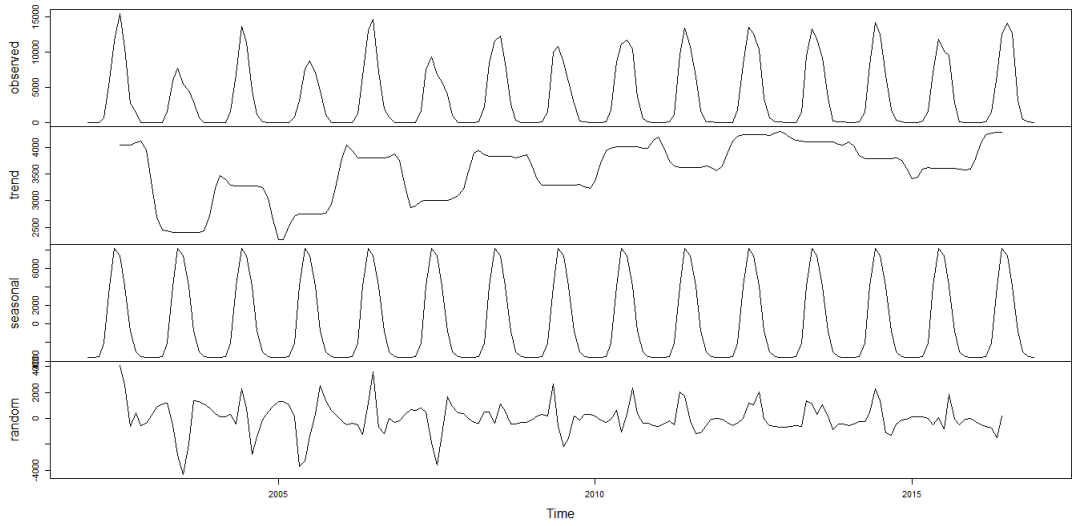
Além disso, o mês de junho é o mês que, historicamente, possui a maior produção seguido por julho e por fim agosto. E também, o mês de Maio e Setembro possuem uma larga escala de produção.

Historicamente, o ano de 2016 é o ano que, acumulado, possui a maior produção de café desde 2002 com 51369 mil sacas de café arábico de conilon colhidas, e logo em seguida, o ano de 2012, 2013 e 2002 foram anos de maior produção para o Brasil, com o total acumulado ao longo dos doze meses de, respectivamente, 50826,4 mil, 49151 mil e 48480 mil sacas de café colhidos.

É interessante notar que após a alta produção de café em 2002, o ano de 2003 foi o pior ano da série, com apenas 28820 mil sacas colhidas e que no ano seguinte a produção foi de 39272 mil sacas.

Além disso, é possível observar a sazonalidade e tendência no gráfico com a decomposição da série abaixo:

Figura 13 – Decomposição da Produção de Café no Brasil (em milhões de sacas) desde 2002



Ao observar a figura acima é possível notar a sazonalidade constante ao longo dos anos e, além disso, pode-se observar que existe uma tendência crescente mais acentuada a partir de 2011.

5.1.1 Transformação do Banco de Dados

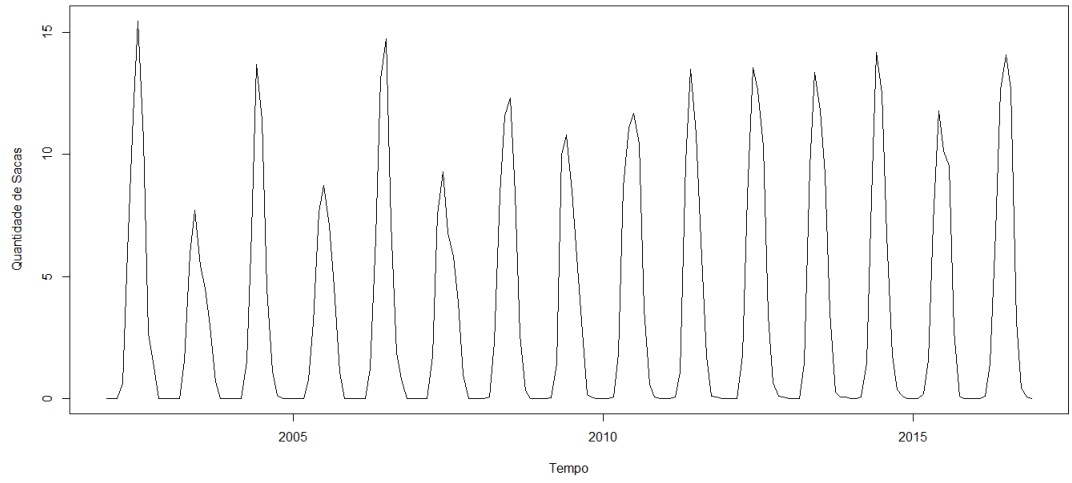
A flexibilidade do Modelo Dinâmico Linear permite que diversos modelos podem ser combinados com outros modelos a fim de gerar um modelo mais generalizado, ou seja, assim como será feito, pode-se separar dois modelos: um com tendência e outro com sazonalidade.

Serão feitas duas transformações no banco de dados a fim de adequa-los a estrutura de Modelagem Dinâmica Linear.

Primeiramente, pela sazonalidade, o banco de dados acima contém ocorrências de produção nulas nos meses de novembro, dezembro, janeiro e fevereiro. Porém, o Brasil possui produtores regionais que sustentam o consumo interno no país, assim, foi admitido para as datas que não possuem produção, a fim de adequar o banco de dados, serão admitidas a produção de 0,0001 mil sacas para compensar as pequenas produções e simplificar a modelagem e transformar o dados matematicamente bem definidos nos reais. E isso foi utilizado para contornarmos o problema de muitos valores iguais a zero que prodem prejudicar a análise.

Nas análises acima, nós trabalhamos com a produção de café no Brasil "em Mil Sacas", porém, a fim de contabilizarmos também as pequenas produções, a partir de agora, as análises serão feitas "Em Milhões de Sacas". Assim, observe o gráfico de produção abaixo:

Figura 14 – Produção de Café no Brasil (em milhões de sacas) desde 2002



Observa-se que, visualmente, o acréscimo da produção regional nos meses de novembro, dezembro, janeiro e fevereiro não foram significativas e, como nas análises acima, percebe-se que temos uma série com sazonalidade implícita e tendência leve.

Além do mais, vamos considerar o banco de dados atual uma série logaritmica

Y_{180x1}.

Definiremos a série y_t como:

$$y_t^* = ln(y_t)$$

onde $t = 1, 2, ..., 180$, sendo como cada um dos meses da série descrita começando em janero de 2002 até dezembro de 2016. Além disso, a priori estamos trabalhando com a série sob suposição de normalidade.

5.2 Modelagem Dinâmica na Produção de Café com 12 períodos

Como estamos lidando com modelos dinâmicos lineares, iremos considerar ao modelo final como a soma de uma série com tendência linear e sazonalidade. Todas as previsões descritas nesse trabalho são para o ano de 2017.

A fim de determinar as melhores variâncias por AdHoc, foram selecionados valores nos quais criasse um equilibrio entre a distribuição a priori e a distribuição observada.

Os parâmetros iniciais do modelo de tendência linear (que consideraremos um modelo de ordem 2) são iguais aos valores estimados por um modelo de regressão.

Os parâmetros do modelo são:

- $F = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & ... & 0 \end{bmatrix}'_{13\times1}$
- $V = \frac{1}{1000}$
- $G = \begin{bmatrix} 1 & 1 & .. & 0 & & .. & 0 & \\ 0 & 1 & & & & & & \\ 0 & 0 & -1 & -1 & .. & -1 & -1 & -1 \\ 0 & & 1 & 0 & .. & & & 0 \\ 0 & & .. & 1 & 0 & & .. & 0 \\ 0 & & 0 & & 1 & .. & & .. \\ 0 & & 0 & 0 & .. & 1 & 0 & 0 \\ 0 & & 0 & 0 & .. & 0 & 1 & 0 \end{bmatrix}_{(13\times13)}$

- $W = \begin{bmatrix} 10^{-4} & 0 & .. & 0 & & .. & 0 & \\ 0 & 10^{-4} & & & & & & \\ 0 & 0 & 0 & 0 & .. & 0 & 0 & 0 \\ 0 & & 0 & 0 & .. & & & 0 \\ 0 & & .. & 0 & 0 & & .. & 0 \\ 0 & & 0 & & 0 & .. & & .. \\ 0 & & 0 & 0 & .. & 0 & 0 & 0 \\ 0 & & 0 & 0 & .. & 0 & 0 & 0 \end{bmatrix}_{(13\times13)}$

- $m_0 = \begin{bmatrix} 0,1519 \\ -0.0003 \\ 0.1454 \\ 1.4060 \\ 0.7403 \\ -0.9454 \\ -1.4141 \\ -0.7832 \\ 0.0526 \\ 0.6211 \\ 1.6157 \\ 0.0573 \\ -1.4958 \end{bmatrix}$

- $C_0 = 0.003738889 \times I_{13x13}$

O vetor m_0 fornece dados da tendência linear, que representam o valor esperado no mês inicial e crescimento esperado (mesmo que baixo) a cada mês, e valores sazonais, que se alternam periodicamente, em um período de 12 meses.

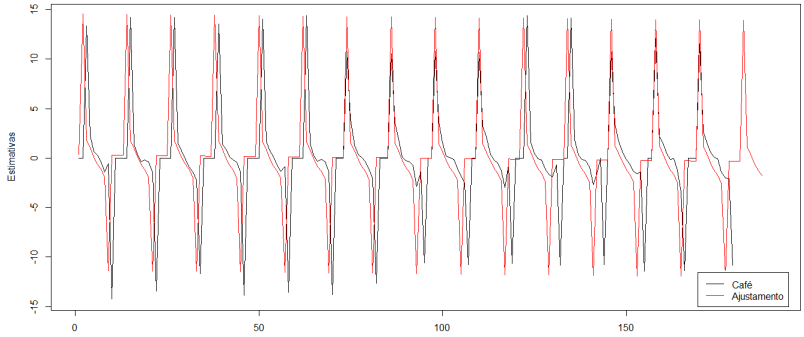
Como visto, a matriz C_0 tem grande relevância em simulações, porém, sabe-se que a mesma C_0 representa a variância a priori dos parâmetros do sistema e o principal ponto de alteração da mesma é ao calcularmos o intervalo de confiança de previsão no qual distingue a escala dos intervalos.

A tratarmos das variâncias V e W, sabe-se que a razão entre as mesma determina se a aproximação do modelo está mais próximo de uma tendência geral ou dos próprios dados observados, ou seja, o critério de seleção dos mesmos é a combinação que representasse melhor o comportamento dos dados.

Valores mais altos e discrepantes de V indicam que as estimativas se aproximam da tendência dada pelo modelo, enquanto que, as estimativas são mais próximas dos dados observados ao colocarmos valores mais altos de W.

Abaixo observamos a série de previsões da série logaritmica escolhida:

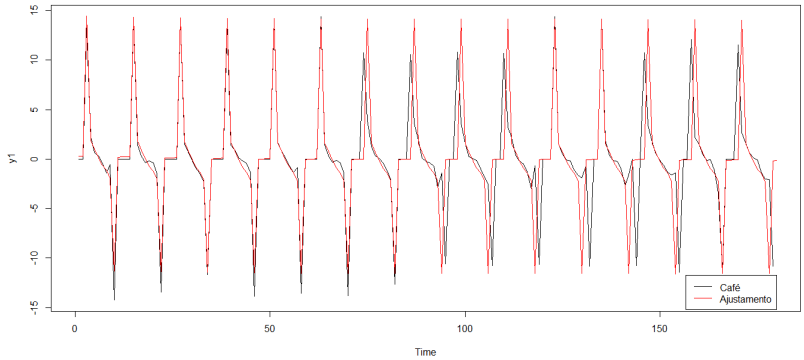
Figura 15 – Modelo Preditivo - Produção de Café no Brasil



Ao observar a previsão acima, nota-se que os valores próximos ao longo da série são muito próximos dos valores adiante. Porém, por se tratar de uma produção em milhões de sacas, por mais que a curva se assemelhe aos dados reais, esses valores encontram-se um pouco distante da escala da produção de 2017.

Agora vamos observar a série suavizada utilizando todos os valores observados ao longo do tempo.

Figura 16 – Modelo Suavizado - Produção de Café no Brasil



A sazonalidade expressiva ao longo da série prejudica os resultados no alisamento do modelo, porém, ainda sim, é possível observar não existem momentos com grandes variações no período e, conforme o tempo passa, a tendência da mesma permanece constante transformando a série em uma série comportada.

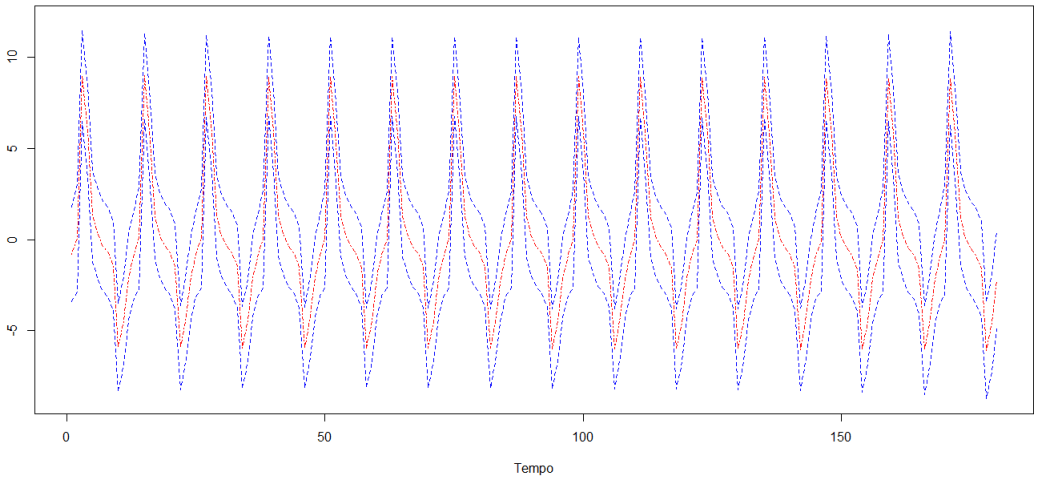
Como esperado, as estimativas para os valores da série são viesados, porém, pode-se avaliar a consistência do mesmo por meio de uma álgebra matricial, calcula-se a variância dos nossos estimadores. O cálculo é feito da seguinte forma:

$$var(\widehat{y}_t) = F_t^{'} \times U_t \times D_t^2 \times U_t^{'} \times F_{t'}$$

onde $U_{(n\times n)}$ é uma matriz ortogonal e $D_{(n\times n)}$ é uma matriz diagonal.

Vamos criar um intervalo de credibilidade de 95% fim de verificar a consistência das nossas estimativas.

Figura 17 – Modelo Ajustado com intervalo de credibilidade de 95% - Produção de Café no Brasil



Por conta da sazonalidade expressiva da série, é possível observar que não é possível identificar um percepção explícita sobre o crescimento ou decrescimento da série de café.

5.2.1 Modelos Alternativos para a série com 12 periodos

Esta seção irá tratar sobre outras alternativas de modelos para a produção de café no Brasil, como por exemplo o modelo ARIMA (médias móveis, auto regressivo). Os modelos a serem analisados são os: ARIMA (5,1,4), ARMA(9,3), SARIMA(1,0,3)(2,1,2), SARIMA(1,1,3)(1,0,1) por esses modelos possuírem uma sazonalidade variável e tendência linear.

Os modelos a serem testados são modelos frequentistas, assim, será possível observar a diferença entre a ténica bayesiana e as frequentistas.

O critério para poder avaliar a eficácia do modelo será pelo Erro Quadrático Médio, que podemos calcular por meio da fórmula:

$$EQ = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2$$

Observe abaixo o gráfico dos modelos ARIMA(5,1,4), ARMA(9,3), SARIMA(1,0,3)(2,1,2), SARIMA(1,1,3)(1,0,1).

Figura 18 – Modelo ARIMA(5,1,4) - Produção de Café no Brasil

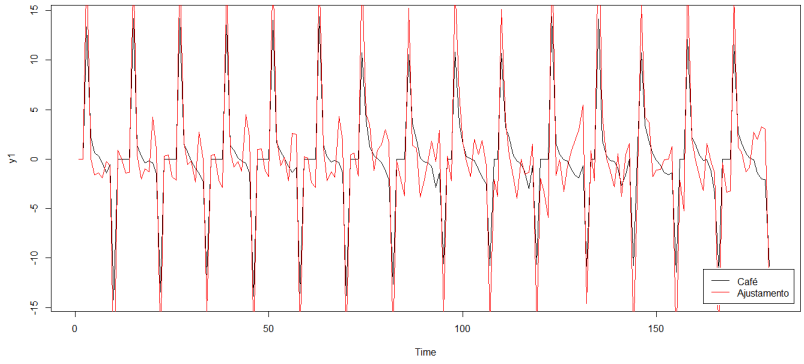


Figura 19 – Modelo ARMA(9,3) - Produção de Café no Brasil

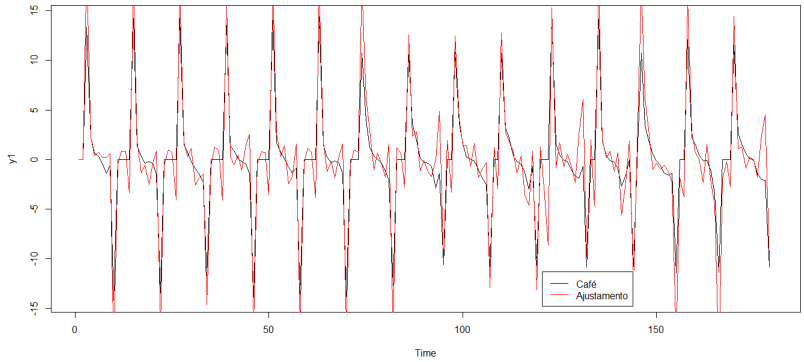


Figura 20 – Modelo SARIMA(1,0,3)(2,1,2) - Produção de Café no Brasil

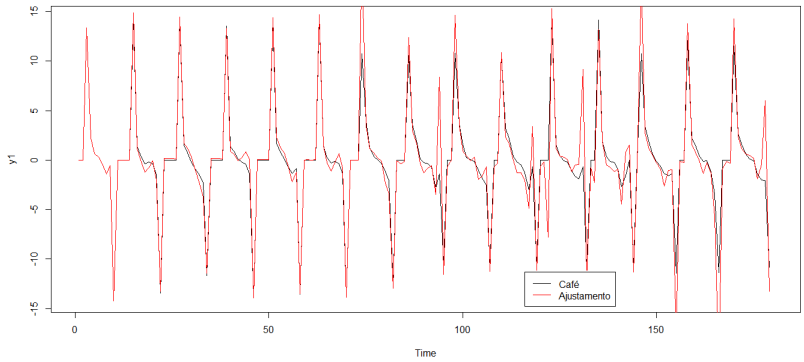
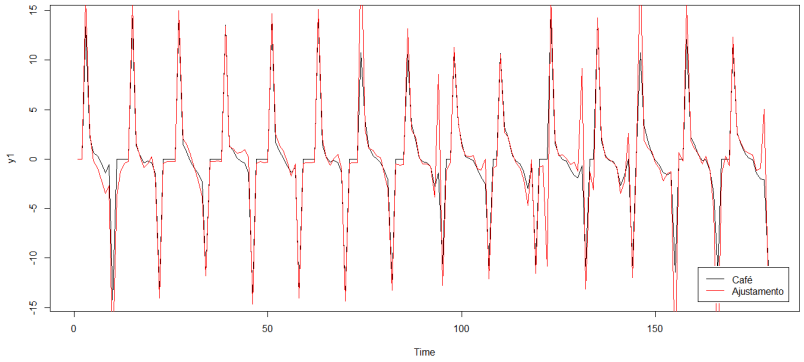


Figura 21 – Modelo SARIMA(1,1,3)(1,0,1) - Produção de Café no Brasil



Ao observar os gráficos acima, nota-se que nenhuma modelo se adequa perfeitamente a série por conta da sazonalidade muito destacada, porém, nota-se que os modelos SARIMA(1,0,3)(2,1,2) e o modelo SARIMA(2,1,3),(1,0,1) possuem uma discrepância menor que os outros modelos.

Agora, observe os erros quadráticos médios de cada um dos modelos abaixo:

Tabela 2 – Erros Quadráticos Médios - Produção de Café no Brasil

Modelo	**Erro Quadrático Médio**
DLM	341.9034
ARIMA(5,1,4)	326.7575
ARMA(9,0,3)	194.2404
SARIMA(1,0,3)(2,1,2)	131.7409
SARIMA(2,1,3),(1,0,1)	157.0448

Ao observar os Erros Quadráticos Médios acima, observa-se que nenhum modelo possuí um valor de Erro Quadrático baixo, ou seja, nenhum dos modelos acima se adequa perfeitamente a série descrita, assim, prejudicando uma previsão de qualidade em todos os casos.

Modelos com uma soma de erros quadráticos menores são mais precisos. As diferenças entres os EQM são muito distantes, assim, pode-se supor que o modelo dinâmico linear não possuiu a melhor adequação.

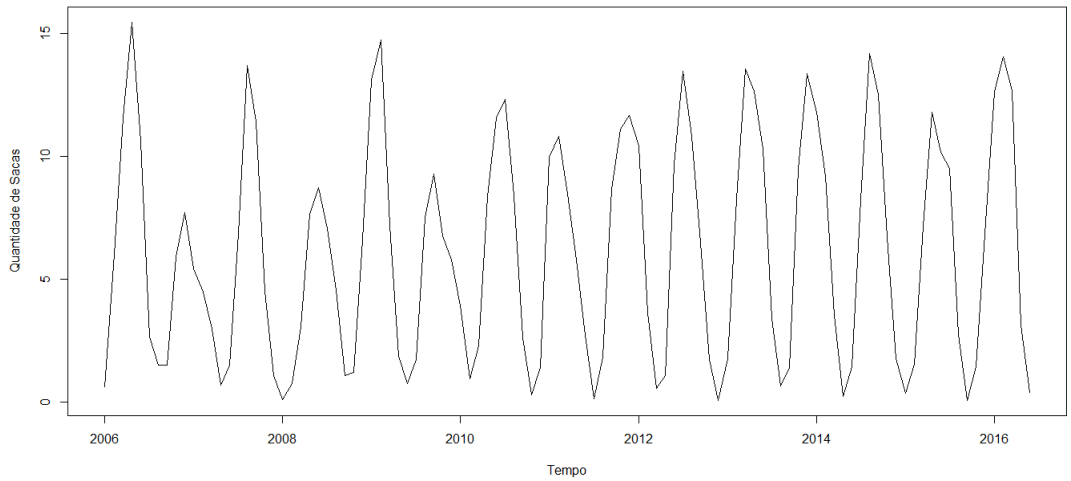
5.3 Modelagem Dinâmica na Produção de Café com 7 periodos

Ao observar a modelagem dinâmica no período completo acima, observa-se que os resultados não são precisos, assim, optou-se por uma nova alternativa. Assim como no modelo de 12 periodos, todas as previsões descritas nesse trabalho são para o ano de 2017.

Como citado acima, os meses Novembro, Dezembro, Janeiro, Fevereiro e Março possuem produção muito baixa, podendo estar atrapalhando a qualidade da análise em questão. Assim, a partir deste ponto, serão desconsiderados os meses acima, consequentemente, teremos uma série com período igual a 7, ou seja, contendo apenas o meses de Abril até Outubro.

Observe o gráfico da nova série abaixo:

Figura 22 – Produção de Café no Brasil de Abril a Outubro



Será utilizado uma série sazonal com 7 períodos, ou seja, ao utilizarmos a soma dos 7 valores sazonais teremos 6 graus de liberdade. Agora, podemos notar que não perdemos a sazonalidade clara da série, ou seja, é possível aplicar a metodologia estudada nos dados.

Observe os parâmetros selecionados abaixo:

- $F = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & ... & 0 \end{bmatrix}'_{8\times 1}$
- $V = \frac{1}{400}$
- $G = \begin{bmatrix} 1 & 1 & .. & 0 & & .. & 0 & \\ 0 & 1 & & & & & & \\ 0 & 0 & -1 & -1 & .. & -1 & -1 & -1 \\ 0 & & 1 & 0 & .. & & & 0 \\ 0 & & .. & 1 & 0 & & .. & 0 \\ 0 & & 0 & & 1 & .. & & .. \\ 0 & & 0 & 0 & .. & 1 & 0 & 0 \\ 0 & & 0 & 0 & .. & 0 & 1 & 0 \end{bmatrix}_{(8\times 8)}$

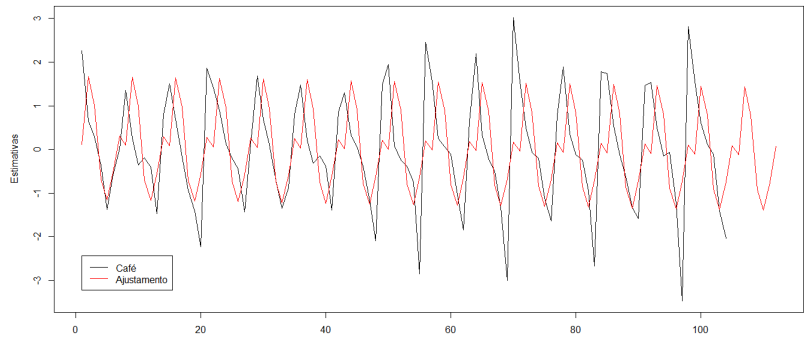
- $W = \begin{bmatrix} 10^{-3} & 0 & .. & 0 & & .. & 0 & \\ 0 & 10^{-3} & & & & & & \\ 0 & 0 & 0 & 0 & .. & 0 & 0 & 0 \\ 0 & & 0 & 0 & .. & & & 0 \\ 0 & & .. & 0 & 0 & & .. & 0 \\ 0 & & 0 & & 0 & .. & & .. \\ 0 & & 0 & 0 & .. & 0 & 0 & 0 \\ 0 & & 0 & 0 & .. & 0 & 0 & 0 \end{bmatrix}_{(8\times 8)}$

- $m_0 = \begin{bmatrix} 0.11257 \\ -0.00224 \\ 0.209968 \\ -0.62596 \\ -1.2568 \\ -0.78817 \\ 0.89769 \end{bmatrix}$

- $C_0 = 345.7143 \times I_{8x8}$

As análises dos parâmetros são elaboradas como na seção anterior, assim, vamos observar o gráfico com o modelo preditivo da nova série com o novo período abaixo:

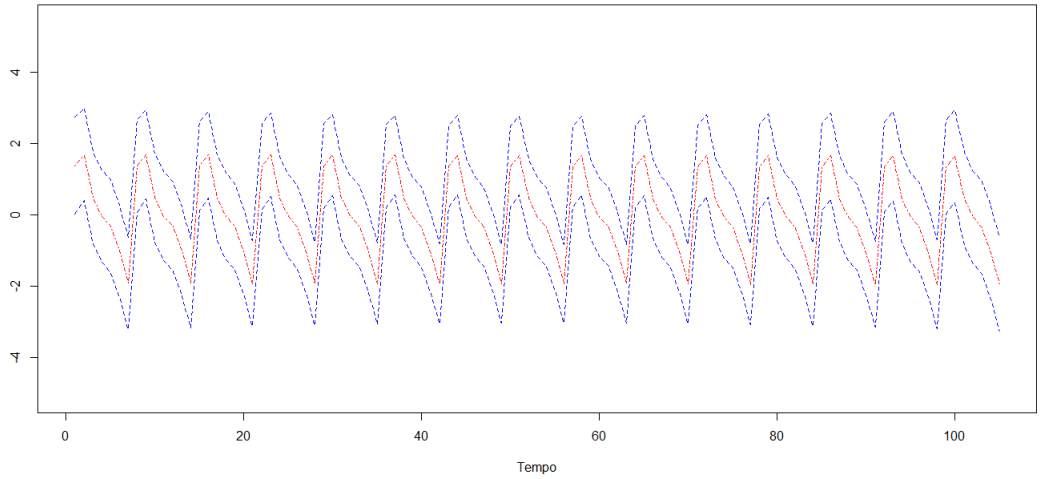
Figura 23 – Modelo Preditivo - Produção de Abril até Outubro de Café no Brasil



Ao observar o gráfico acima, nota-se que a previsão está mais condizente com uma possível previsão mais precisa da série.

Seguindo o mesmo princípio de seção acima, observe o gráfico abaixo com o intervalo de credibilidade da nova série abaixo:

Figura 24 – Modelo Prediivo - Produção de Abril até Outubro de Café no Brasil de Abril a Outubro



Diferentemente da primeira série, possível observar que o intervalo de confiança está seguindo toda a série, ou seja, existem evidências para acreditar que a priori desta série possui mais informação que a anterior.

5.3.1 Modelos Alternativos para a série com 7 periodos

Seguindo os mesmos princípios, nesta seção iremos avaliar e comparar outras alternativas de modelos para as estimações para a nova série de 7 períodos.

Observe abaixo o gráfico dos modelos ARIMA(5,1,4), ARMA(9,3), SARIMA(1,0,3)(2,0,2), SARIMA(1,1,3)(1,0,1) para os meses de Abril até Outubro.

Figura 25 – Modelo ARIMA(5,1,4) - Produção de Café no Brasil de Abril a Outubro

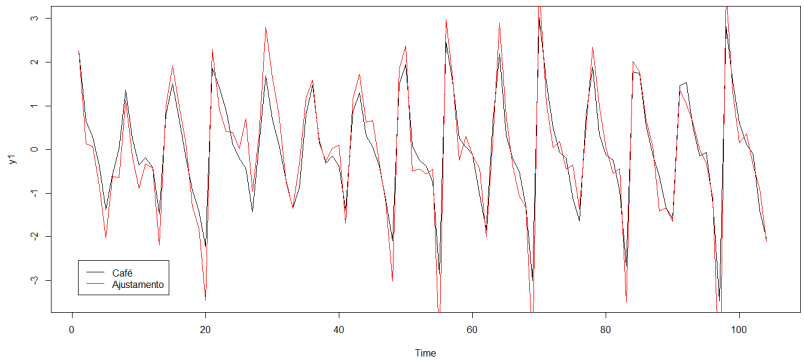


Figura 26 – Modelo ARMA(9,3) - Produção de Café no Brasil de Abril a Outubro

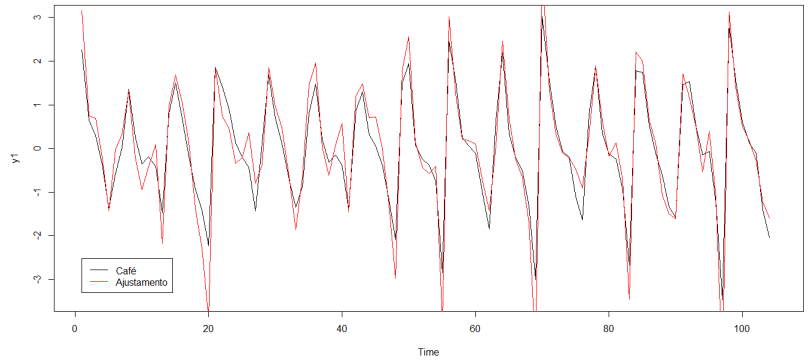


Figura 27 – Modelo SARIMA(1,0,3)(2,0,2) - Produção de Café no Brasil de Abril a Outubro

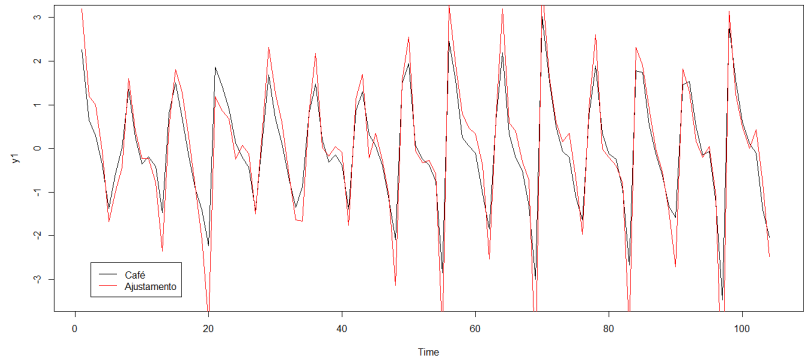
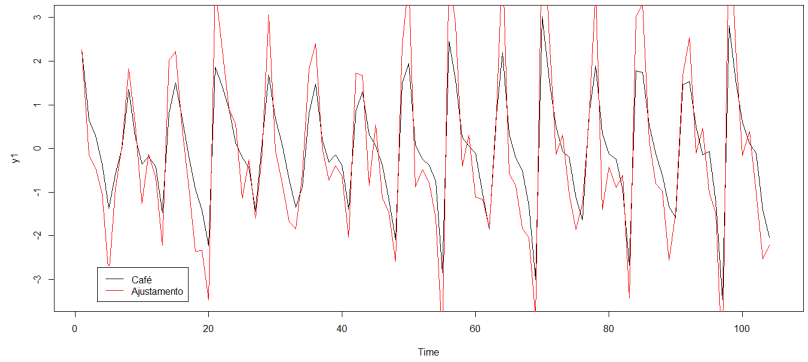


Figura 28 – Modelo SARIMA(1,1,3)(1,0,1) - Produção de Café no Brasil de Abril a Outubro



Ao comparar os gráficos acima, primeiramente, os modelos desta série, em geral, estão melhor adequados a série em questão que a série anterior e, além disso, nota-se que os modelos ARIMA(5,1,4) e o SARIMA(1,0,3)(2,1,2) são os modelos que possuem, visualmente, melhor adequação a série de produção de café no Brasil.

Agora vamos observar os Erros Quadráticos Médios de cada uma das séries alternativas:

Tabela 3 – Erros Quadráticos Médios para Modelos Alternativos - Produção de Café no Brasil pro ano de 2017

Modelo	**Erro Quadrático Médio**
ARIMA(5,1,4)	30.2385
ARMA(9,0,3)	93.8072
SARIMA(1,0,3)(2,0,2)	37.2408
SARIMA(2,1,3),(1,0,1)	97.5986

Por hora, a fim de obter uma estimação ainda mais precisa, foram testadas diversas combinações das mastrizes de variâncias a fim de obter o Erro Quadrático do modelo dinâmico linear mais baixo possível, observe a tabela com alguns resultados abaixo:

Tabela 4 – Erros Quadráticos Médios para DLM - Produção de Café no Brasil pro ano de 2017

Variâncias	**Erro Quadrático Médio**
V= 1000 e W = 20000	155, 0291
V = 10 e W = 400	133,8501
V = 1/400 e W = 1/1000	126,0585
V<1/2000 e W<1/50000	104,0612

Como dito acima, quanto menor o Erro Quadrático Médio, melhor ajustada é a série, para a previsão do ano de 2017, assim como vimos no gráfico, os erros quadráticos médios desta série são melhores que os da série anterior e, além disso, pode-se dizer que o modelo ARIMA(5,1,4) e o SARIMA(1,0,3)(2,1,2) realmente ajustam-se melhor a série de produção de café no Brasil.

Já olhando para os EQM'S dos Modelos dinâmicos, nota-se que a partir de, aproximadamente, $V = 1/2000$ e $W = 1/50000$, o erro quadrático não sofreu alteração significante a ponto de selecionarmos outra opção de modelo.

E assim a fim de verificar e validar os ajustes de cada modelo acima, foi calculado a log-verossimilhança de cada um dos deles, observe a tabela com os resultados abaixo:

Tabela 5 – Log Verossimilhança para Modelos Alternativos - Produção de Café no Brasil pro ano de 2017

Modelo	**Log Verossimilhança**
DLM	-179.1053
ARIMA(5,1,4)	-123.2397
ARMA(9,0,3)	-108.2825
SARIMA(1,0,3)(2,0,2)	-139.9897
SARIMA(2,1,3),(1,0,1)	-156.5175

Por meio da função *dlmLL*, calculou-se a log verossimilhança do modelo dinâmicos linear e notou-se que o valor se aproximou dos valores de log verossimilhança dos outros

modelos, assim, permitindo a aceitação da utilização da metodologia para os dados.

6 Conclusão

Segundo a CONAB, no ano de 2016, foram produzidos aproximadamente 51,6 milhões de sacas no Brasil e no ano de 2017, a produção de café no Brasil foi de 44,97 milhões de sacas, ou seja, decaiu 12,5% com relação a produção absoluta de 2016.

Ao observar os valores previstos para o ano de 2017 pela Modelagem Dinâmica na série de 12 períodos, observa-se que houve um decrescimento de aproximadamente 30% comparado com a produção de 2016 da CONAB, ou seja, houve uma divergência de aproximadamente 18,5% entre a realidade e a modelagem Bayesiana.

Ao analisar a previsão par ao ano de 2017, foi possível comparar esse percentual de produção de café com a série com 7 períodos pois, para a previsão desse caso, foi suposto que o meses entre Novembro e Março são nulos e, ao analisar os dados, observou-se que a produção prevista é aproximadamente 12% que a real, ou seja, ao comparar a produção real absoluta de 2017 com a produção prevista pela modelagem, houve uma dispersão de 12%

Mesmo que possam existir outras opções de modelo mais precisos, como o SARIMA(1,0,3)(2,1,2), e mesmo a Modelagem Dinâmica não dando a melhor previsão possível para os dados, com as transformações adequadas, foi possível utiliza-la no banco de dados a fim de obter estimações viáveis para a produção de café no Brasil no ano de 2017. Porém, graficamente, pode-se notar que o modelo dinâmico linear conseguiu captar bem o comportamento da série.

O resultado encontrado é razoável pois foi possível explorar grande parte da metodologia de Modelos Dinâmicos Lineares aplicado no banco de dados da produção café no Brasil, acreditamos, que com o acréscimo de mais variáveis explicativa seria possível melhorar o comportamento de tendência e sazonalidade da série.

Além disso, acreditamos que seria possível melhorar a estimação pois as matrizes de variância poderiam ser estimadas por meio da metodologia bayesiana como o Amostrador de Gibbs.

E assim, acreditamos que este trabalho possa ser uma motivação para trabalhos futuros utilizarem Modelagem Dinâmica Linear em dados reais, que é uma técnica bayesiana consideravelmente recente.

Referências Bibliográficas

Campagnoli, P., Petrone, S., Petris, G. (2009). Dynamic Linear Models with R.

Chopin, Nicolas. "Dynamic detection of change points in long time series."Annals of the Institute of Statistical Mathematics 59.2 (2007): 349-366.

Correia, Leandro Tavares (2010). Modelos dinâmicos para dados agregados.

Fan, Xueping, et al. "Mathematical simulation of coupled fluid flow and geomechanical behavior for full low permeability gas reservoir fracturing [J]."Petroleum Exploration and Development 27.1 (2000): 76-83.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. Jornal of Applied Statistics, 10:1–18.

Gamerman, D. Lopes, H. F. (2006). Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press.

McCullagh, P. and Nelder, J. (1994). Generalized Linear Models, volume 37. Monographs on Statistics and Applied Probability, 2nd edition.

Petris, G. (2010). An R package for dynamic linear models. Journal of Statistical Software, 36(12):1–16.

Petris, G., Petrone, S., Campagnoli, P. (2009). Dynamic linear models. Dynamic Linear Models with R, pages 31–84.

Rodrigues, Guilherme Souza. Modelos dinâmicos Dirichlet.

Sevestre, P. Trognon, A. (1996). Dynamic linear models. In: The Econometrics of Panel Data, pages 120–144. Springer.

Silva, P. H. D. d. (2016). Modelos dinâmicos com pontos de mudança para dados de contagem.

West, M. and Harrison, P. (1997). Bayesian Forcast and Dynamic Models. Springer Verlag, 2nd edition

West, M. (1996). Bayesian forecasting. Wiley Online Library.

West, M., Harrison, P. J., Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting. Journal of the American Statistical Association, 80(389):73–83.

Ziegel, E. R. (1997). Bayesian forecasting and dynamic models. Technometrics, 39(4):433.